

CMS

CERN

LHC

[Large Hadron Collider]

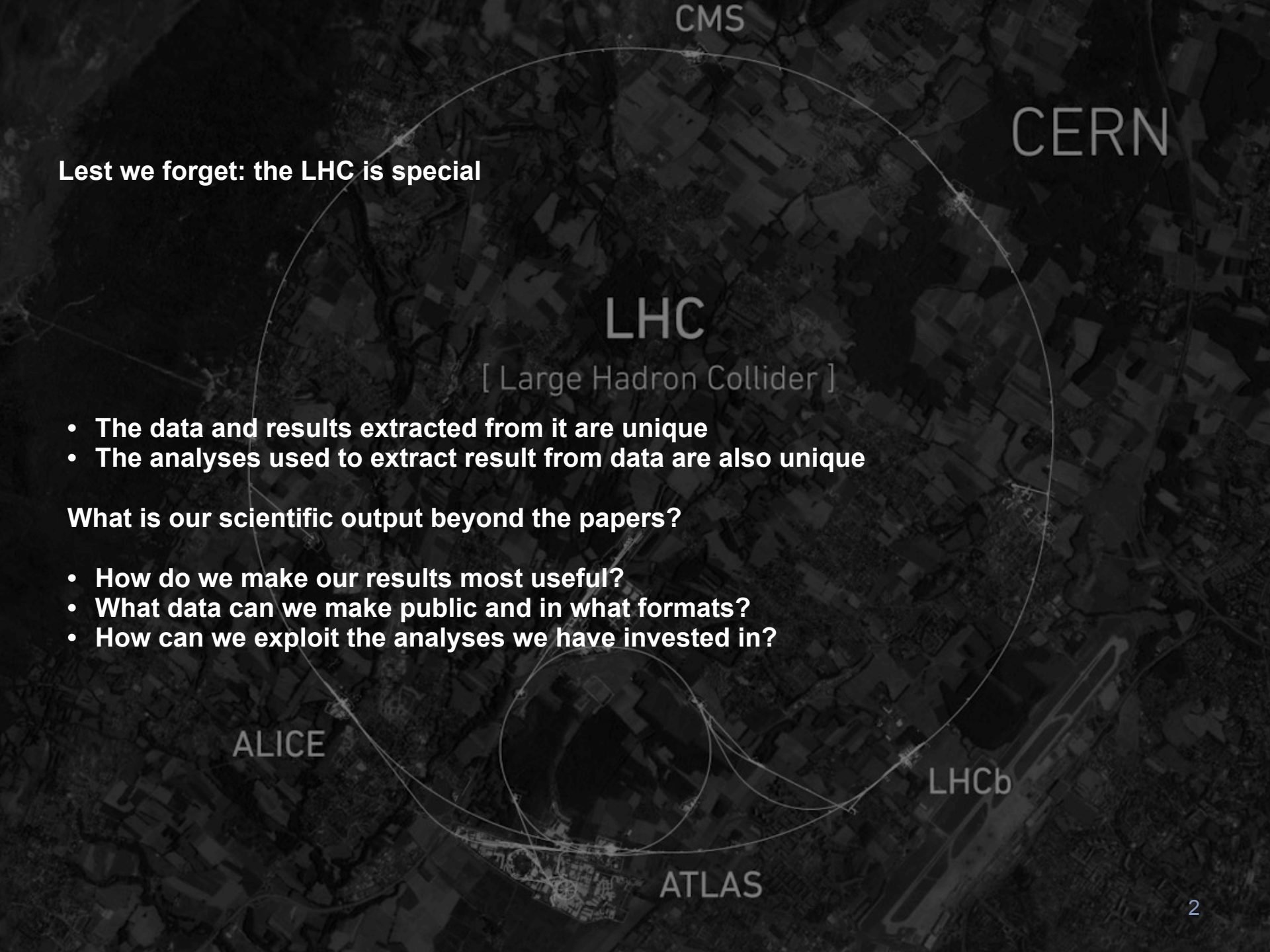
Analysis Preservation Open Data, RECAST

ALICE

L Heinrich
Snowmass 2020

LHCb

ATLAS

An aerial photograph of the CERN facility in Geneva, Switzerland, with a dark, high-contrast filter. A large white circle outlines the LHC tunnel. Inside this circle, the text 'LHC' is prominently displayed in the center, with '[Large Hadron Collider]' written below it in a smaller font. Four smaller white circles are also visible, each representing a different experiment: CMS at the top, ALICE at the bottom-left, ATLAS at the bottom, and LHCb at the bottom-right. The labels for these experiments are placed near their respective circles. The word 'CERN' is written in large, white, sans-serif capital letters in the upper right quadrant of the image.

Lest we forget: the LHC is special

- **The data and results extracted from it are unique**
- **The analyses used to extract result from data are also unique**

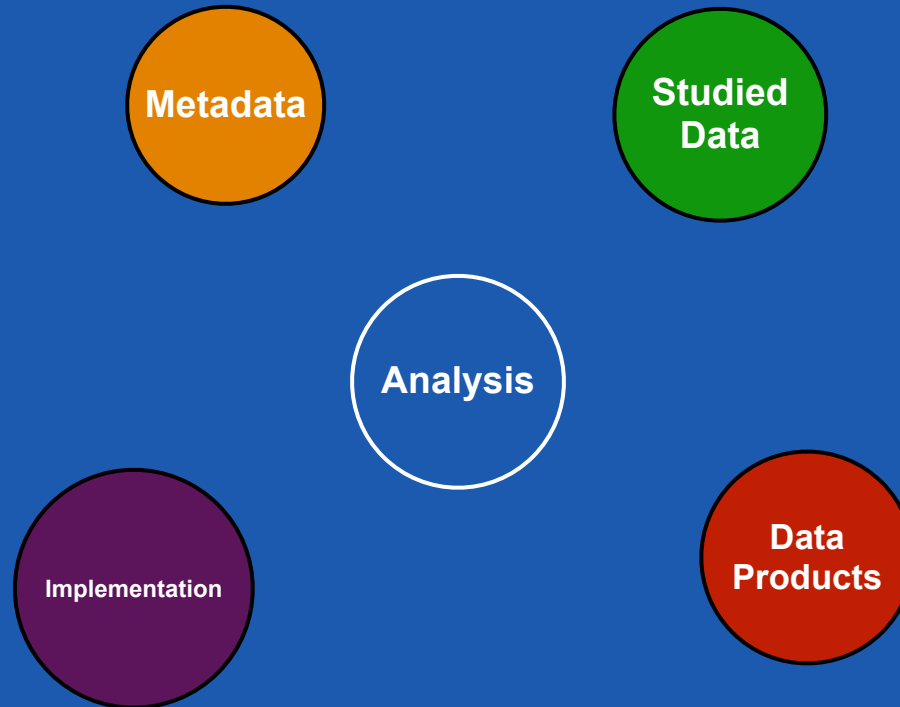
What is our scientific output beyond the papers?

- **How do we make our results most useful?**
- **What data can we make public and in what formats?**
- **How can we - as Collaborations - exploit the analyses we have invested in?**

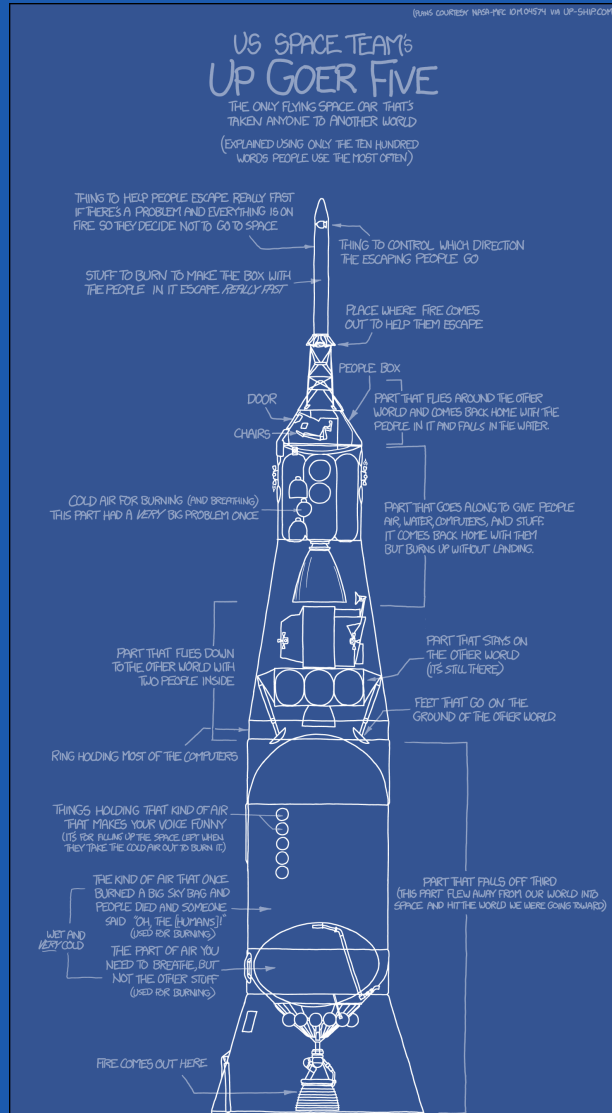
Analysis Sketch



Preservation Domains



External

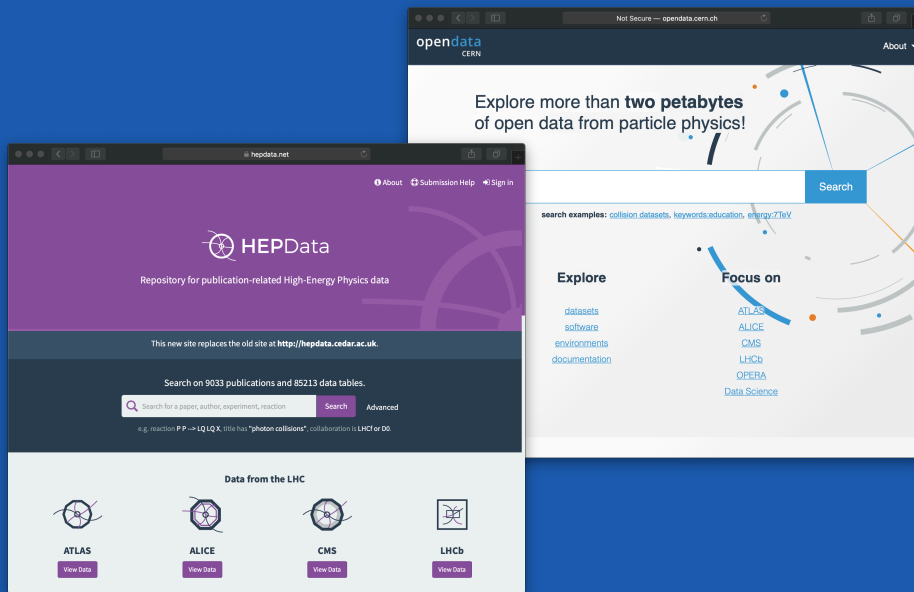


Internal



Three broad areas of activity

External



Analysis Data Products
and Result Preservation

Open Data for Outreach,
Education and Research

Internal



Reproducible Workflows &
Analysis Preservation

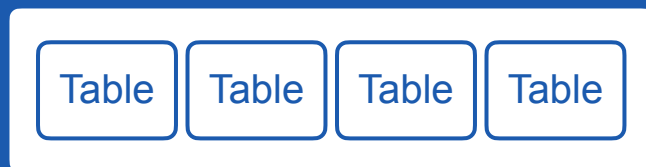
HepData has been the main vehicle to provide

high quality public **data products** for published analyses

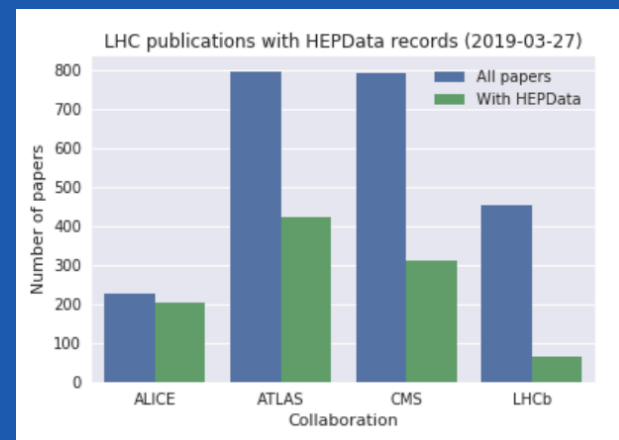
publicly available. All LHC experiments rely on this.

- HepData submission often required for analysis approval

Types of data products expanded from



to broader collection of data



ALICE: 90%
ATLAS: 52%
CMS: 39%
LHCb: 14%

Additional Material helps approximate reimplementation of data analyses w/ e.g. Rivet (can cover also BSM and HI)

```
#include "SimpleAnalysis/AnalysisClass.h"
#include "SimpleAnalysis/NtupleMaker.h"
#include "SimpleAnalysis/PDFReweight.h"
#include <LHAPDF/LHAPDF.h>
#include "TMath.h"

DefineAnalysis(EwkOneLeptonTwoBjets2018)
// Wh->l+bb+met analysis (Run2 data)

void EwkOneLeptonTwoBjets2018::Init()

{
    // Define signal/control regions
    // Define signal/control regions
```

```
// -*- C++ -*-
#include "Rivet/Analysis.hh"
#include "Rivet/Projections/ChargedFinalState.hh"
#include "Rivet/Tools/Correlators.hh"
#include "Rivet/Tools/AliceCommon.hh"
#include "Rivet/Projections/AliceCommon.hh"

namespace Rivet {

    /// @brief Multiparticle azimuthal correlations pp, pPb, XeXe and PbPb
    class ALICE_2019_I1723697 : public CumulantAnalysis {
    public:

        /// Constructor
        ALICE_2019_I1723697() :
            CumulantAnalysis("ALICE 2019 I1723697") {}
    };
}
```

Confronting Experimental Data with Heavy-Ion Models

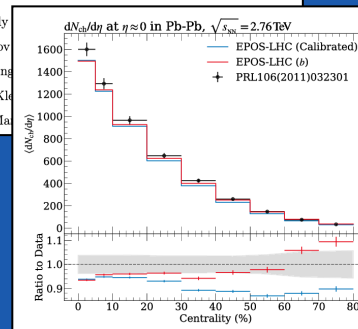
RIVET for Heavy Ions

Christian Bierlich,^{1,2} Andy

Peter Harald Lindenov

Jan Fiete Grosse-Oetring

Patrick Kirchgaesser,⁶ Jochen KleChristine O. Rasmussen,² Ma

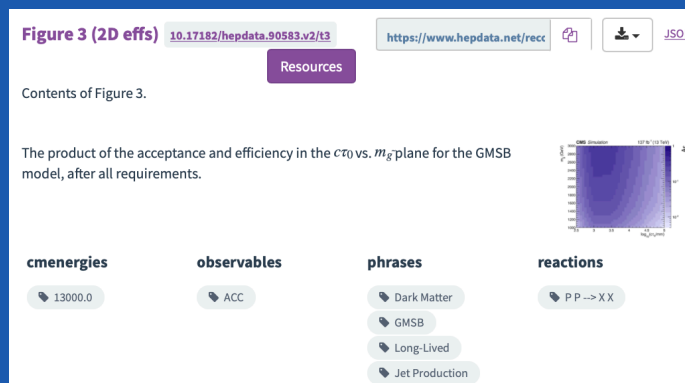


C++ Code Snippets as starting point, or (better) full Rivet Routine

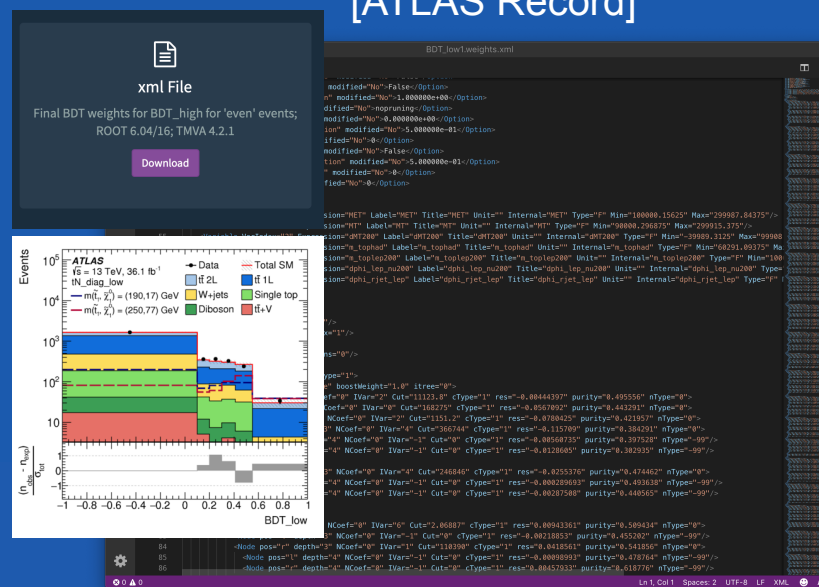
Efficiency Maps:

ML models uploaded to HepData

[ATLAS Record]



[CMS Record]

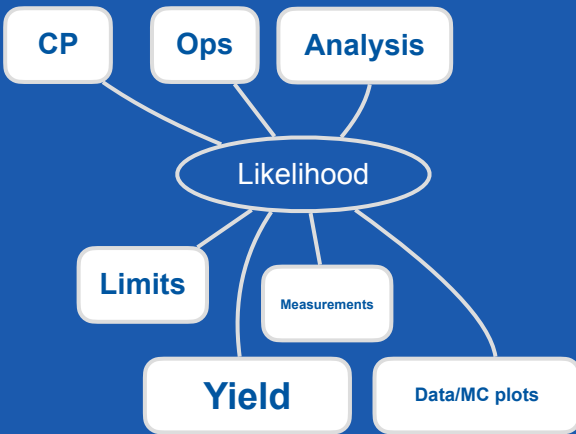


A new Frontier: Public Likelihoods

$$p(\text{theory}|\text{data}) \sim p(\text{data}|\text{theory}) \cdot p(\text{theory})$$

likelihood: experimentalists

prior: theorists



The likelihood is the central object in analysis

- the best data product we can provide in principle

Often HepData information (yields, uncertainties...) is used to reconstruct approximate likelihood



2000

2010

2012

2017

2019

HepData is for experimentalists interacting with wider community by releasing public information about already existing data analyses.

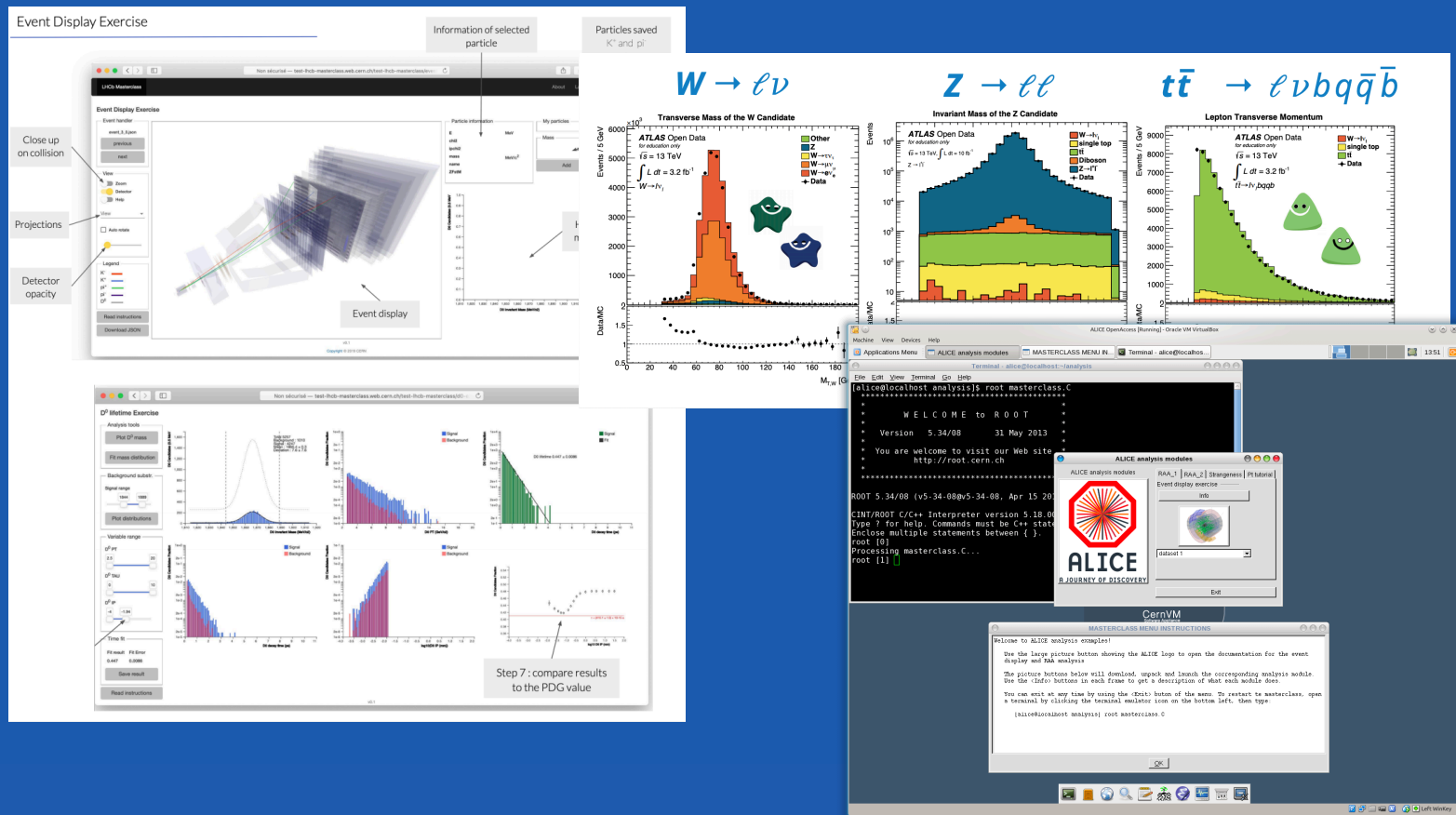
But it **may not be enough to interact with community on developments of new analyses techniques.**

For this we might require a more free-form mode of collaboration: Open Data

Open Data

All LHC Experiments have Open Data Programs

- integrated into CERN Open Data Portal
- ATLAS, LHCb, ALICE so far focused mainly on Outreach & Education



CMS has more expansive Open Data Program for Research

We see external eco-system developing

- Workshop in October [link]

Number of Papers appearing on e.g.
Machine Learning methods for LHC

Exploring the Space of Jets with CMS Open Data

Patrick T. Komiske^{1,2,*}, Radha Mastandrea^{1,1}, Eric M. Metodiev^{1,2,1}, Preksha Naik^{1,1} and Jesse Thaler^{1,2,1}

¹Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
²Department of Physics, Harvard University, Cambridge, MA 02138, USA

End-to-end particle and event identification at the Large Hadron Collider with CMS Open Data

M. Andrews¹, J. Alison¹, S. An^{1,2}, P. Brvant¹, B. Burkle³, S. Glevzer⁴, M. Narain³, M. Paulini¹, B.

¹Department
³Departme
⁴Departme
⁵Machine Learn

Fast and accurate simulation of particle detectors using generative adversarial networks

Pasquale Musella · Francesco Pandolfi

26 Nov 2018

Abstract Deep generative models parametrised by neural networks have recently started to provide accurate results in modeling natural images. In particular, generative adversarial networks provide an unsupervised solution to this problem. In this work we apply this

the date of receipt and acceptance should be inserted later

Noname manuscript No.
(will be inserted by the editor)

arxiv:1908.08542

arxiv:1910.07029

arxiv:1805.00850

EnergyFlow

Search docs

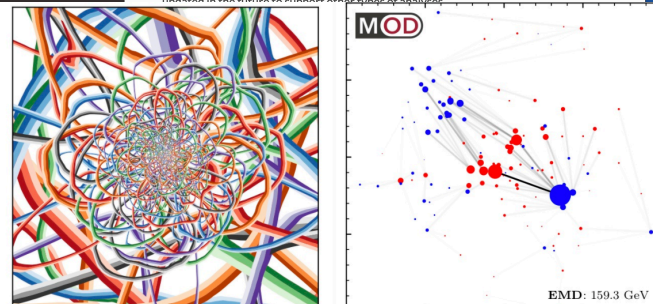
Home
Getting Started
Installation
Demos
Examples
FAQs
Release Notes
News
Documentation
Architectures
Datasets
CMS Open I
HDF5 Form
V000000
GitHub

Docs » Documentation » Datasets

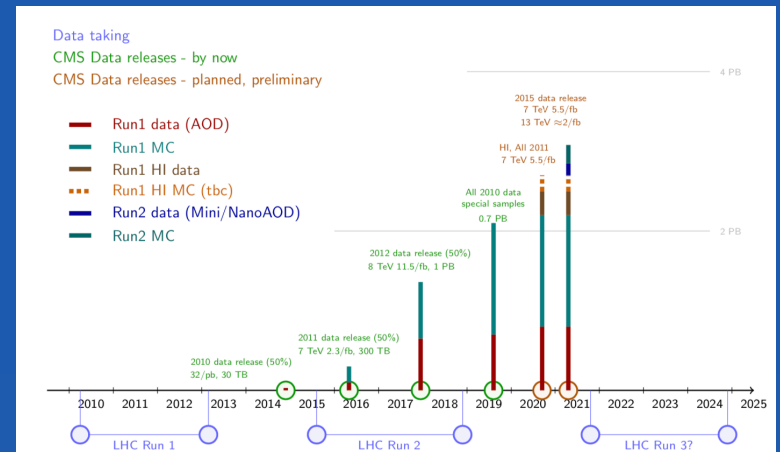
CMS Open Data and the MOD HDF5 Format

Starting in 2014, the CMS Collaboration began to release research-grade recorded and simulated datasets on the [CERN Open Data Portal](#). These fantastic resources provide a unique opportunity for researchers with diverse connections to experimental particle physics world to engage with cutting edge particle physics by developing tools and testing novel strategies on actual LHC data. Our goal in making portions of the CMS Open Data available in a reprocessed format is to ease as best as possible the technical complications that have thus far been present when attempting to use Open Data (see also [recent efforts by the CMS Collaboration](#) to make the data more accessible).

To facilitate access to Open Data, we have developed a format utilizing the widespread [HDF5 file format](#) that stores essential information for some particle physics analyses. This "MOD HDF5 Format" is currently optimized for studies based on jets, but may be extended in the future to encompass other physics analyses.

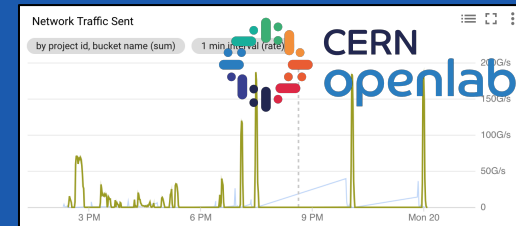


Release Schedule being finalized for coming years



Currently we see a range in approaches. Questions raised by Open Data (L3):

- Can we ensure ability to perform analysis of sufficient quality
- What data formats / software would be released? What's the level of support (if any?)
 - how "final" should objects be (ability to re-reconstruct...)
- Protect Collaboration / Cohesion
 - without a strong collaboration preparation of high-quality OD impossible
- Is analyzing PB scale data really feasible for external users.
 - emerging public cloud infrastructure might help provide on-demand access to scale
 - maybe targeted datasets (e.g. for ML R&D) rather than blanket Open Data?



KubeCon 2019 Keynote [link]

CERN is seeking to harmonize them via a common Open Data Working Group.

HepData is for experimentalists interacting with wider community by releasing public information about already existing data analyses.

Open Data might be useful for experimentalists to interact with external researches on new R&D

But both approaches are not enough.

**HepData is analysis-specific, but lossy.
Open Data for new work outside of expt's**

→ need infrastructure for internal, lossless analysis preservation

Internal analysis preservation can capture detail unavailable in other modes of data/analysis preservation.

Increasing complexity in analyses to fully exploit potential of LHC dataset

- **low-level observables: (e.g. BDT on calorimeter clusters)**
 - even if we publish BDT / NN weights, can you reliably simulate those low-level details?
- **whole-event observables (NN inputs from many objects)**
 - simple description of signal model acceptance via e.g. efficiency tables will not work anymore
- **Need a way to preserve analyses part of result pipeline at full fidelity.**



Internal Reuse:

Efforts by all LHC experiments to foster internal analysis preservation.
Ingredients for AP:

capture software

archive analysis code incl.
dependencies

capture commands

what do with the
captured software

capture workflow

order of individual steps

data assets

input data needed
to run the analysis

CERN provides infrastructure to
assist experiments

REANA: workflows-as-a-service

CAP: store workflow and
other analysis artifacts
(software, etc)



reana

Reproducible research data analysis platform

CERN
Analysis Preservation

capture, preserve and reuse physics analyses



1. capture software

archive analysis
code incl. deps.

2. capture commands

what do with the
captured software

3. capture workflow

order of individual
steps

Containers universally seen as suitable technology:

all experiments have some
infrastructure to run
experiment / analysis code
in containers

REANA Environment AliPhysics

build unknown glitter join chat License GPL v2

About

`reana-env-aliphysics` provides a container image with encapsulated runtime execution environment for AliPhysics based ALICE data analyses. The container image includes all the necessary dependencies and does not have any external requirements (such as CVMFS).

`reana-env-aliphysics` was developed for use in the REANA reusable research data analysis platform.

lhcb-analysis-preservation > containerization-cookie > Details

C

containerization-cookie

Project ID: 31307 | [Leave project](#)

45 Commits 1 Branch 0 Tags 287 KB Files

Cookiecutter template for analysis containerization



atlas/athanalysis

By atlas • Updated 8 days ago

ATLAS Athena Analysis Release

Container

1M+ Downloads 3 Stars



atlas/analysisbase

By atlas • Updated 8 days ago

ATLAS Standalone Analysis Release

Container

1M+ Downloads 10 Stars

clelange / cmssw-docker

<> Code

Issues 5

Pull requests 0

Actions

Projects 0

Dockerfiles for CMSSW <https://doi.org/10.5281/zenodo.3374807>

82 commits

1 branch

0 packages

Tag: v1.0

New pull request



clelange Use --build-arg instead of wrong -e for docker ENV

1. capture software

archive analysis
code incl. deps.

2. capture commands

what do with the
captured software

3. capture workflow

order of individual
steps

Workflow languages seem to be a good choice:

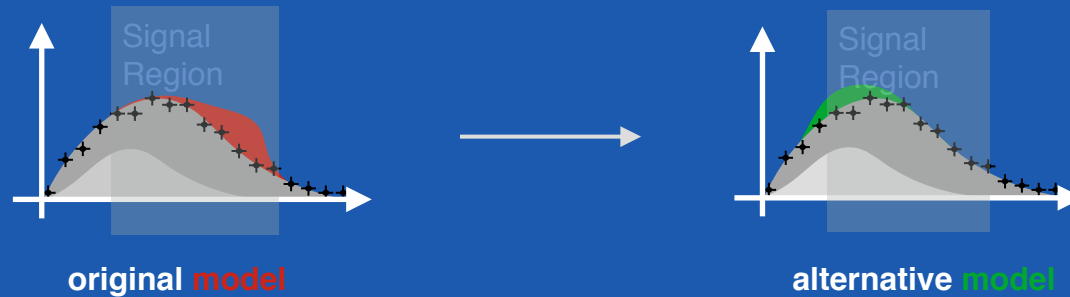
REANA supports Common Workflow Language, Yadage

- looking into snakemake (LHCb also has snakemake starter kit)

The image displays three overlapping screenshots of the REANA GitHub repository, each showing a different analysis example. The top-left screenshot shows the 'REANA example - ALICE LEGO train test run' page, which includes a 'build passing' badge and a 'license: GPL-2.0' badge. The top-right screenshot shows the 'REANA example - CMS Higgs-to-four-leptons' page, which includes a 'build: unknown' badge and a 'license: MIT' badge. The bottom-center screenshot shows the 'REANA example - ATLAS RECAST' page, which includes a 'build: passing' badge and a 'license: MIT' badge. Each page has an 'About' section and an 'Analysis structure' section. The 'About' section for the ATLAS RECAST example includes a mathematical equation:
$$D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^-$$
. The 'Analysis structure' section for the ATLAS RECAST example includes a list of inputs: '1. Input data' and 'The analysis takes the following inputs:'. The bottom-left screenshot shows a terminal window with the following commands:

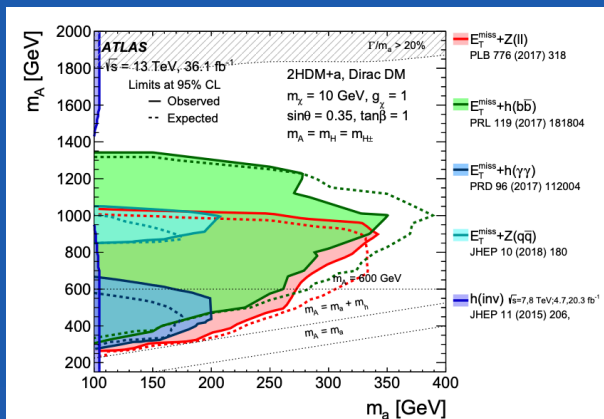
```
$ mkdir -p _alice_data_2010_LHC10h_2_000139038
$ cd _alice_data_2010_LHC10h_2_000139038
$ wget http://opendata.cern.ch/record/1102/files/assets/alice/2010/LHC10h/000139038/ESD/000139038.root
$ cd ..
```

Major use-case for internal re-use: reinterpretation

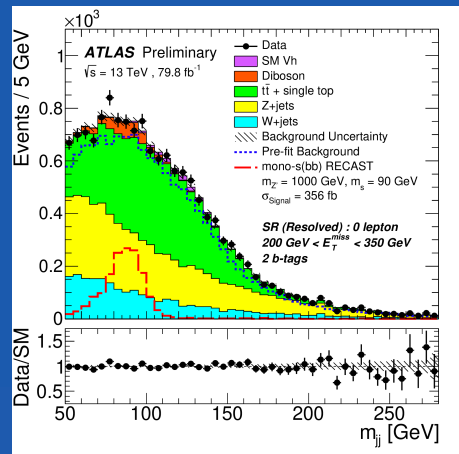


ATLAS: require analyzers to preserve analysis that at least reinterpretation w/ REANA is possible → realization of RECAST (docker images, scripts, workflows)

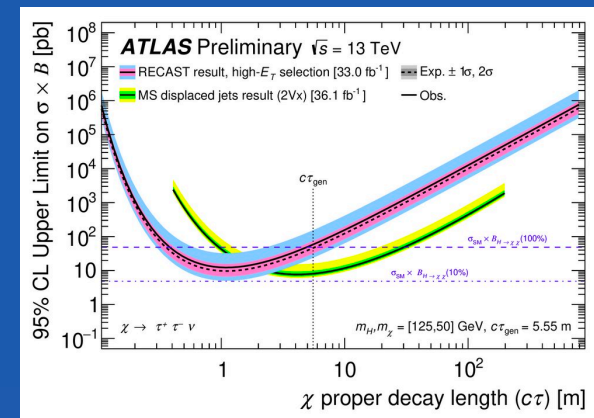
New scientific results based on this (rather technical) requirement



arxiv:1903.01400



ATL-PHYS-PUB-2019-032



ATL-PHYS-PUB-2020-007



CERN Analysis Preservation Examples

CERN Analysis Preservation

BETA

Web

Files | Data | Source Code

LBZLcDOK.tur.gz (419 KB)

...

BASIC INFORMATION

ANALYSIS NAME

MEASUREMENT

PROPOSERS

NAME

NAME

ORCID

NAME

NAME

NAME

STATUS

INSTITUTES INVOLVED

KEYWORDS

STRIPPING/TURBO SELECTIONS

TYPE OF DATASET

CUSTOM NAME

STRIPPING/TURBO LINE

BOOKKEEPING LOCATIONS

TYPE OF DATASET

CUSTOM NAME

STRIPPING/TURBO LINE

BOOKKEEPING LOCATIONS

NTUPLE/USERDEF-PRODUCTION

CUSTOM NAME

INPUT DATASET

PLATFORM

DAVINCI VERSION

OUTPUT EOS LOCATION

LBZLcDOK branching fraction

LBZLcDOK branching fraction

BR(LbLcD0aK⁰) and BR(LbLcD0⁰aK⁰) with respect to LBZLcS⁰

Sebastian Neubert

Hartun Tschal

0900-0001-8476-8188

Alexsio Pflüci

Nicola Sedmore

1 - in preparation

8.2

Pentapequarks

ms_data

data 2012

XZ(LbD0D0XNPBeauty2CharmLine

/J/psiKaCollison12/Bran04005GeV-VetoClosed_McNlo/Real Data/Rec14/Stripping21/1900000000/SHADRON.MDST

/J/psiKaCollison12/Bran04005GeV-VetoClosed_McNlo/Down/Real Data/Rec14/Stripping21/1900000000/SHADRON.MDST

ms_data

data 2011

XZ(LbD0D0XNPBeauty2CharmLine

/J/psiKaCollison11/Bran13005GeV-VetoClosed_McNlo/Real Data/Rec14/Stripping21/1700000000/SHADRON.MDST

/J/psiKaCollison11/Bran13005GeV-VetoClosed_McNlo/Real Data/Rec14/Stripping21/1700000000/SHADRON.MDST

2011 samples

data 2011

u65_64-u65-qc62-opt

u4x5

/eos/tnb/wg/Bar04/EXOTIC/LbLcDOK/Impiles/Stripping21

USER ANALYSIS

Copyright 2018 © CERN. Created & Hosted by CERN. Powered by Invenio Software.

[Contact](#)
[About](#)
[Search Tips](#)

About Search Tips

LHCb

16 results		< > Page 1 of 2
SUBJECTS		
<input type="checkbox"/> draft	16	JME-10-004
<input checked="" type="checkbox"/> TYPE		
<input type="checkbox"/> cms-analysis-v0.0.1	16	MUON
PHYSICS_OBJECTS		
<input checked="" type="checkbox"/> jet	22	FWD-10-005
<input checked="" type="checkbox"/> muon	16	
<input type="checkbox"/> PFMuon	10	MUON
<input type="checkbox"/> GlobalMuon	4	
<input type="checkbox"/> TrackerMuon	4	
<input type="checkbox"/> electron	10	AN-2011/103
<input type="checkbox"/> photon	6	
<input type="checkbox"/> MET	2	ELECTRON MUON
<input type="checkbox"/> tau	2	
<input type="checkbox"/> track	2	
<input type="checkbox"/> vertex	2	AN-2011/062
		MUON
		AN-2011/103
		ELECTRON MUON
		AN-2010/411
		ELECTRON MUON MET

CMS

Focus on:

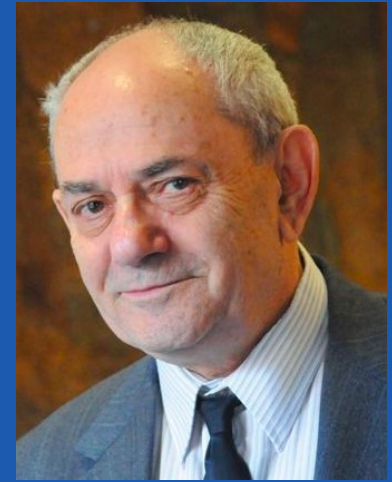
- ease of use for analysis teams
 - e.g. auto-complete, automatic ingestion, command line clients
- ease of use for users
 - discoverability / search
 - integration with REANA



We see a pattern:

Likelihood: Preserving general likelihood functions in a sustainable way is a hard problem.

→ restricting to binned models (HistFactory) enabled progress by narrowing scope



"When solving a problem of interest, do not solve a more general problem as an intermediate step"
- V. Vapnik

AP: Preserving Analyses in full generality (does it work in 100 years?) is too big a problem. RECAST focuses on

- **near-/mid-term solution (e.g. assume containers)**
- **subset of the problem of reinterpretation, not e.g. re-estimation of background)**

We seem to be entering a golden era of data / analysis preservation for LHC.

Absence of new physics (so far!) forces us to focus on full exploitation of data.

We finally have the technology:

- **containers / workflows for software/analysis preservation**
- **at-scale compute on-demand for Open Data**

See some sociological shifts in community:

- **likelihood releases**
- **analysis preservation (RECAST) as approval requirement**

Good opportunity for new strategic efforts for both open and internal preservation efforts.

